## NCI Center for Bioinformatics

# Informatics Seminar Series

**9:00 until Noon**
**June 23, 2003**
**6116 Executive Blvd., Conf. Room 3056A/B**

—

## From Protein Sequence to Function:

### Functional Analysis of Protein sequences and Protein Classification

**(9:00 until 10:30)**

**Anastasia Nikolskaya**
**Protein Information Resource**
**Department of Biochemistry and Molecular Biology**
**Georgetown University Medical Center**

—

## Extraction of Protein Names from MEDLINE:

### An Evaluation Perspective

**(10:30 until noon)**

**Inderjeet Mani**
**Senior Principal Scientist**
**The MITRE Corporation**

—

Following each presentation there will be an open discussion.

**Notes for Dr. Nikolskaya's presentation**

The availability of complete genomes has dramatically changed modern biology by finally making it possible to catalogue all proteins that are responsible for every essential cellular function and to identify functions that are missing in each particular organism. However, the accumulation of genome sequences poses a challenge of predicting the functions of proteins that are not yet experimentally characterized.

Objectives of functional analysis for different groups of protein sequences will be discussed: (1) functional assignment/annotation of protein sequences of "known" function, common sources of mistakes and how to avoid them; (2) functional prediction for the more complex cases, some tools, methods and hints.

Functional annotation of individual sequences could be greatly aided by comprehensive and fully annotated protein classification system(s). Three protein classification databases will be discussed: Protein Information resource (PIR, http://pir.georgetown.edu/ ), InterPro (http://www.ebi.ac.uk/interpro/ ), and Clusters of Orthologous Groups of proteins (COGs) (http://www.ncbi.nlm.nih.gov/COG/). Classification databases are important tools for extracting information from genome sequences for functional prediction/annotation, biochemical pathway analysis, and evolutionary genome analysis.

## Notes for Dr. **Mani's** presentation

Given the vast amounts of genomic and molecular data being generated by scientific research, there is a pressing need to develop advanced bioinformatics infrastructures for biological knowledge management. Ontologies for biology are crucial in data integration from multiple databases and in literature mining for knowledge extraction and evidence attribution. Ontology development, however, currently requires substantial human effort. This talk provides an overview of a research project at Georgetown University aimed at exploiting computational linguistics tools to induce ontology of protein names using text corpora from MEDLINE. The talk goes on to discuss methods, evaluation issues, and results in protein name extraction.

Inderjeet Mani is a Senior Principal Scientist at MITRE and an adjunct faculty in computational linguistics at Georgetown University. He has been a Principal Investigator on projects in summarization, temporal information extraction, and information retrieval funded by NSF, DARPA, and MITRE. His publications include several books and numerous papers.